第27卷 第6期 工 程 数 学 学 报 Vol. 27 No. 6
2010年12月 CHINESE JOURNAL OF ENGINEERING MATHEMATICS Dec. 2010

文章编号:1005-3085(2010)06-0995-06

基于形状信息的Bayes分类方法*

王 妍[1,3]， 刘 君[1]， 柴阿丽[2]， 黄海洋[1,†]， 郇中丹[1]， 李宝聚[2]

(1- 北京师范大学数学科学学院 数学与复杂系统教育部重点实验室，北京 100875;
2- 中国农业科学院蔬菜花卉研究所，北京 100081; 3- 北京财贸职业学院，北京 101101)

摘 要: 本文提出了一种新的基于形状信息的Bayes分类方法，以实现对图像中单个物体的分类。该方法首先运用图像边缘提取和配准算法，构造一个形状相似性能量泛函，并利用其计算形状信息的先验概率。然后，结合图像中物体其它特征的后验概率，通过Bayes方法进行分类。本文将该方法应用于一个病原菌图像分类的实际问题，实验结果表明，该方法是十分有效的，不仅降低了分类所需的特征维数，而且提高了分类精度，并能满足实际问题中所要求的计算速度。

关键词: Bayes分类；形状信息；图像配准；单个物体

分类号: **AMS**(2000) 68T10 **中图分类号**: O23; O175 **文献标识码**: A

1 引言

图像分类是图像理解、图像检索等领域的基本问题，在视频监控、卫星图像、医学图像等实际场合都有比较广泛的应用；同时由于图像本身的大数据量特点，研究图像的分类对于模式识别的研究也具有重要的理论意义，涉及到图像处理与计算机图像学研究中的很多基本方面，例如特征提取[1,2]、形状相似性度量[3,4]等。

图像分类中的单一物体图像分类问题，即每幅样本图像中只存在单一待分类物体情况下的分类问题，是图像分类中一类常见的问题，并且具有非常广泛的应用。例如植物病斑[5]、水果[6]、鞋底花纹[7]、肿瘤[8]、血细胞[9]等物体图像的分类问题。在单物体图像分类问题中，物体的形状信息是图像的一个重要特征和分类的重要依据，但形状信息具有难于描述，数据维数较高的特点，给分类带来了一定困难。

Bayes方法[10]是模式识别的基本方法之一，在输入数据满足假设的概率分布的情况下，Bayes分类器是在最小错误率条件下的最优分类器，但是Bayes方法的概率分布比较难于估计，尤其是当输入数据特征维数较高的情况下，还会给分类过程本身带来额外的计算复杂度。

本文提出了一种基于形状信息的Bayes分类方法，以克服以上提到的困难，达到简化计算复杂度和提高分类精度的要求。并将该方法运用于一个病原菌图像分类的实际问题，实验结果表明，与其它不考虑形状的目标分类方法相比较，本文提出的算法可以得到更好的分类精度，同时计算速度基本相同，满足病原菌分类识别的实用要求。

收稿日期: 2009-03-05. **作者简介**: 王妍(1984年3月生)，女，硕士. 研究方向：图像处理，模式识别，数学建模.
*基金项目: 国家自然科学基金(10531040)；国家“863”项目(2006AA10Z210).
†通讯作者: 黄海洋 E-mail: hhywsg@bnu.edu.cn

2 基于形状信息的贝叶斯分类方法

在多类识别问题中，假设有n类样本，设表示n个类别的集合为：$\Omega = \{X_1, X_2, \cdots, X_n\}$。设测试样本为$x$，由适当的特征信息表示。欲判断测试样本$x$属于$\Omega$中的哪一类，最好基于测试样本特征信息的概率来确定。记$p(x \in X_i \,|\, x)$为在已知测试样本x的条件下，x属于X_i类的条件概率。如果$l = \arg\max_{1\le i\le n} p(x \in X_i \,|\, x)$，就判定$x$属于第$l$类。通常根据Bayes公式计算条件概率，$p(x \in X_i \,|\, x) = p(x \,|\, X_i)p(X_i)/p(x)$，其中$p(x)$表示测试样本$x$出现的概率，根据$x$的特征计算；$p(X_i)$表示$X_i$类出现的概率，一般由专家给出，称为先验概率；$p(x \,|\, X_i)$是在$X_i$类的情况下，$x$出现的概率，一般要通过训练样本特征来估计，称作后验概率。

为克服形状信息提取和处理的难点，本文将形状信息从特征中分离出来处理。定义样本的新变量为$x := (\tilde{x}, \varsigma)$，其中$\tilde{x}$表示测试样本$x$的传统特征信息，如面积周长比、边缘长度等，$\varsigma$表示测试样本$x$的形状信息，并采用修正的Bayes公式

$$p(x \in X_i \,|\, x) = \frac{p(\tilde{x}|(X_i, \varsigma))p(X_i|\varsigma)}{p(\tilde{x}|\varsigma)}. \tag{1}$$

也就是将样本x的分类过程分解为两个步骤：先求在已知样本x形状特征ς前提下，X_i类出现的先验概率$p(X_i \,|\, \varsigma)$。严格的说，这一项不是传统Bayes意义下类别的先验概率，而是加入了测试样本的形状特征，称为“形状先验概率”，然后求在已知X_i类别和形状特征ς前提下，样本x传统特征$\tilde{x}$出现的后验概率$p(\tilde{x} \,|\, (X_i, \varsigma))$，根据修正的Bayes公式得到分类结果。

后验概率$p(\tilde{x} \,|\, (X_i, \varsigma))$相对容易计算，例如，针对病原菌分类的实际特点，我们提取传统的特征$\tilde{x} = (\tilde{x}_1, \tilde{x}_2, \tilde{x}_3) : \tilde{x}_1$为边缘长度$l$，$\tilde{x}_2 = (4\pi S)/(l^2)$为面积$S$周长$l$比，$\tilde{x}_3$为归一化的灰度分布，即病原菌内部灰度值落在$[0, g_b - \sigma_L]$，$(g_b - \sigma_L, g_b + \sigma_H]$，$[g_b + \sigma_H, 255]$三个区间的像素点个数的百分比，其中$g_b$是图像中的背景部分的平均灰度，$\sigma_H$和$\sigma_L$是根据病原菌内部灰度值分布情况取定的2个固定的阈值。对各类训练样本传统特征的统计分析表明每种特征都服从高斯分布，又由于特征之间的相对独立性，我们假设传统特征服从联合高斯分布。按极大似然估计，可以由X_i类训练样本传统特征值得到相应高斯分布的均值μ_i与方差σ_i，从而得到后验概率$p(\tilde{x} \,|\, (X_i, \varsigma))$。

3 形状先验概率

为说明形状先验概率$p(X_i \,|\, \varsigma)$的计算方法，我们先研究刻画形状相近的指标。传统方法一般采用近圆性、偏心率、边界点坐标集等形状特征，导致形状信息损失较多，或形状特征维数过高。为此，我们希望找到一个衡量两个形状相似程度的数学量。本文借用文献[11]中做形状相似性配准所用的能量泛函来刻画形状相似度。首先，我们利用形态学方法或者水平集方法对图像进行边缘提取，将需要分类的对象的边界提取出来。然后，将待分类对象的边界与标准形状做图像配准，计算形状相似度。

设待分类对象的边界为闭曲线$\mathcal{S}$，图像区域为Ω，被$\mathcal{S}$包围的前景区域为$R_{\mathcal{S}}$，背景为Ω中除去$R_{\mathcal{S}}$后的区域。曲线$\mathcal{S}$可以用符号距离函数

$$\phi_{\mathcal{S}}(x) = \begin{cases} 0, & x \in \mathcal{S}, \\ +D(x, \mathcal{S}) > 0, & x \in R_{\mathcal{S}}, \\ -D(x, \mathcal{S}) < 0, & x \in \Omega - R_{\mathcal{S}} \end{cases}$$

的水平集 $\{x;\, \phi_{\mathcal{S}}(x)=0\}$ 表示，其中 $D(x,\mathcal{S})=\min_{y\in\mathcal{S}}\{\|x-y\|_2\}$ 表示点 x 到曲线 $\mathcal{S}$ 的距离。由于相同形状物体的大小、摆放的位置和方向可能不同，只有经过刚性变换处理，才能归类。设 $x=(x_1,x_2)$，刚体变换 $\mathcal{A}: x\mapsto \hat{x}$ 定义如下

$$\mathcal{A}(x_1,x_2)=\alpha\begin{pmatrix}\cos\theta & \sin\theta\\ -\sin\theta & \cos\theta\end{pmatrix}\begin{pmatrix}x_1\\ x_2\end{pmatrix}+\begin{pmatrix}\mu\\ \nu\end{pmatrix},$$

其中 α 表示缩放因子，θ 表示旋转角度，μ、ν 表示平移。下面也用符号 $\mathcal{A}$ 代表这4个参数。记曲线 $\mathcal{S}$ 经 $\mathcal{A}$ 变换后的像为 τ。因为

$$D(\hat{x},\tau)=\min_{\hat{y}\in\tau}\|\hat{x}-\hat{y}\|_2=\alpha\min_{y\in\mathcal{S}}\|x-y\|_2=\alpha D(x,\mathcal{S}).$$

所以 $\phi_\tau(\hat{x})=\alpha\phi_{\mathcal{S}}(x)$。

设 γ 为标准形状，为刻画曲线 $\mathcal{S}$ 与 γ 的相似程度，在图像配准中，通常采用能量泛函

$$Q(\mathcal{S},\gamma)=\min_{\mathcal{A}}\int_\Omega\big(\alpha\phi_{\mathcal{S}}(x)-\phi_\gamma(\mathcal{A}(x))\big)^2\mathrm{d}x. \tag{2}$$

也就是说，通过最佳的刚性变换 $\mathcal{A}$，两条曲线 $\mathcal{S}$ 和 γ 的相似程度可以用泛函数值 $Q(\mathcal{S},\gamma)$ 表示。$Q(\mathcal{S},\gamma)$ 值越小，两条曲线越相似。但是，如果在整个图像区域 Ω 上计算泛函 $Q(\mathcal{S},\gamma)$ 表达式中的积分，那么计算量很大。实际上我们只关心区域 Ω 上样本的边界曲线 $\mathcal{S}$。因此，我们取相似性能量泛函为

$$E(\mathcal{S},\gamma)=\min_{\mathcal{A}}\int_\Omega\psi_\beta\big(\phi_{\mathcal{S}}(x)\big)\big(\alpha\phi_{\mathcal{S}}(x)-\phi_\gamma(\mathcal{A}(x))\big)^2\mathrm{d}x, \tag{3}$$

其中

$$\psi_\beta(\phi)=\begin{cases}1, & 若\ |\phi|<\beta,\\ 0, & 否则\end{cases}$$

是一个边界检测函数，其中 β 是控制选取感兴趣区域大小的参数，本文取 $\beta=0.5$ 个像素间距。于是，积分求和过程实际上只在边界附近执行，大大减少了计算量。又由函数 ϕ 的定义可见，在曲线围成区域的中心，ϕ 值较大。这就意味着在衡量两条闭曲线的相似性时，闭曲线所围区域的中心点有着很大的决定权重。因此，在计算形状相似性能量泛函 $E(\mathcal{S},\gamma)$ 前，我们先对两个形状做中心匹配，也为随后求极小元 $\mathcal{A}$ 的迭代算法找到一个较好初值。

直观想像便可知，形状先验概率 $p(X_i\,|\,\varsigma)$ 应当反比于测试样本 x 的形状 ς 与类别 X_i 的形状之间的相似度能量泛函。类别 X_i 的形状由训练样本的形状集组成。如果将测试样本与每个训练样本进行比对，计算量太大。为了减小计算复杂度，我们在每一类别 i 的训练样本集中，利用专家知识选择具有代表性的形状，这里的选择可以是不唯一的。于是，对每一类别 i 选取的形状代表集合 R_i，我们要求 $p(X_i\,|\,\varsigma)\propto E(x,R_i)^{-1}$，其中 $E(x,R_i)$ 表示样本 x 与形状代表集合 R_i 的形状相似性能量泛函，其定义为

$$E(x,R_i)=\min_{\gamma\in R_i}E(\varsigma,\gamma).$$

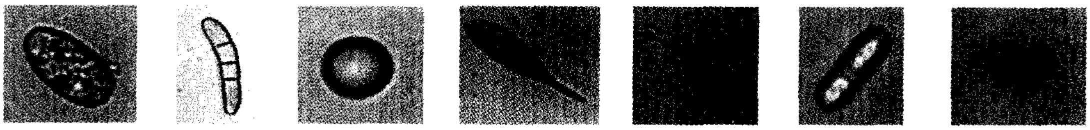
$E(x,R_i)$ 值越小，说明 x 与 X_i 类越接近。所以，在已知 x 的形状 ς 时，X_i 出现的概率 $p(X_i\,|\,\varsigma)$ 越大。再经过归一化处理，我们可以定义形状先验概率为

$$p(X_i\,|\,\varsigma)=\frac{E(x,R_i)^{-1}}{\sum\limits_j E(x,R_j)^{-1}}. \tag{4}$$

这种做法的另一个好处是在对训练样本进行存储时，不必对每个训练样本的形状都进行存储，而只需储存代表形状样本集合中的样本形状即可。一般形状是以图像或是边缘点坐标的形状储存，需要大量的存储空间，因此我们的方法既减小了算法的复杂度，又节省了存储空间。

4 实验结果

为了验证此方法的实验效果，我们对病原菌图像进行分类测试，本节中用到的实验数据均来自实际的病原菌显微镜数据，共包含7类的病原菌图像，其示例图像如图1所示，从实验数据中随机选择一部分作为训练样本，一部分作为测试样本进行实验。在实验中先采用数学形态学方法及水平集方法[12] 进行边缘提取。然后，视具体情况对每类选择1-3个代表形状进行训练。最后用训练得到的基于形状信息的贝叶斯分类器对测试样本进行分类。具体处理情况与分类结果见表1，为了与传统的Bayes分类方法进行对比，表1中同时列出了不引入形状先验概率，直接采用传统特征，运用Bayes方法直接对样本集进行分类得到的正确率结果。



1. 白粉　2. 棒胞　3. 灰霉　4. 链格胞　5. 尾胞　6. 炭疽　7. 叶点霉

图1: 示例病原菌图像

表1: 实验结果

病原菌名称	训练样本数	代表形状数	测试样本数	传统方法正确率	本文方法正确率
1.白粉	10	1	43	83.7%	100%
2.棒胞	14	3	120	66.7%	93.3%
3.灰霉	14	1	27	63.0%	96.3%
4.链格胞	5	2	12	100%	75%
5.尾孢	6	2	17	76.5%	94.1%
6.炭疽	11	2	345	95.1%	81.7%
7.叶点霉	14	1	592	65.4%	91.7%
合计	63	12	1156	75.5%	89.2%

从表中可以看出，我们方法的分类正确率达到了89.2%，与传统的Bayes方法相比正确率有了较明显的提升；另一方面，由于引入了“代表形状”的概念，对于每一个类别，仅需记录训练样本中具有代表性的样本的图像与形状信息，这减小了分类器对存储量的要求。同时，对于每一个测试样本，不需要同每一个训练样本的形状均进行较为繁琐的相似性配准迭代计算，而只需与相对较少的代表形状相对形状比较，注意到代表形状个数与全部训练样本个数之比 $12/63 = 19.0\%$ 仅为大约1/5，代表形状的引入有效的减小了系统分类的计算复杂度，在Matlab环境中，对每个测试样本进行分类平均消耗的时间约为20秒。综上，本系统可以达到并且可以满足实用的分类精度与时间复杂度要求，达到了预期的效果。

5 结论

本文提出了一种基于形状信息的贝叶斯方法，主要是引入形状特征的一种新的表示方法，实现对单一物体图像的分类。采用标准形状本身，使得标准形状信息没有一点损失。通过图像配准，确定样本与标准形状之间的相似度，用相似度的倒数刻画根据形状样本属于各类别的概率，最后采用Bayes方法分类。图像处理方法在这个分类过程中起了重要的作用，不仅如一般图像识别问题一样，需要经过图像去噪、去模糊、增强边缘和提取边缘等过程，而且引用了图像配准方法。这种方法对形状具有明显差异的单一物体图像的分类效果很好，计算速度也很快。但是对多个物体图像的个体识别分类还有困难，特别是，对出现个体间相互重叠的情况，识别效果还很差。尽管我们曾用基于形状先验分割方法提取边缘，但识别的效果不理想，有关这方面的内容还有待于进一步的研究。

致谢：本文中所使用的图像资料及相关生物学知识由中国农业科学院蔬菜花卉研究所提供。

参考文献：

[1] Haralick R M, Shangmugam I, Dinstein K. Textural feature for image classification[J]. IEEE Transactions on System Man and Cybernetics, 1973, SMC-3(6): 610-621

[2] Leu J G. On indexing the periodicity of image textures[J]. Image and Vision Computing, 2001, 19(13): 987-1000

[3] Ang Y H, Li Z, Ong S H. Image retrieval based on multidimensional feature properties[J]. Proceedings of SPIE, 1995, 2420: 47-57

[4] Loncaric S, Dhawan A P. Near-optimal MST-based shape description using genetic algorithm[J]. Pattern Recognition, 1995, (28)4: 571-579

[5] 田有文，张长水，李成华. 支持向量机在植物病斑形状识别中的应用研究[J]. 农业工程学报，2004, 20(3): 134-136
Tian Y W, Zhang C S, Li C H. Application of support vector machine to shape recognition of plant disease spot[J]. Transactions of the Chinese Society of Agricultural Engineering, 2004, 20(3): 134-136

[6] 赵燕超，徐丽明. 基于多元聚类分析的草莓形状分类算法[J]. 中国农业大学学报，2008, 13(1): 77-80
Zhao Y C, Xu L M. Classification approach for shape grading of strawberry based on clustering[J]. Journal of China Agricultural University, 2008, 13(1): 77-80

[7] 管燕，李存华，仲兆满. 一种基于综合特征的鞋底图像识别方法[J]. 西南民族大学学报(自然科学版)，2007, 33(5): 1189-1194
Guan Y, Li C H, Zhong Z M. A recognition method for shoe soles based on integrated features[J]. Journal of Southwest University for Nationalities (Natural Science Edition), 2007, 33(5): 1189-1194

[8] 王彬，孙蕾. 基于支持向量机的肿瘤形状特征分类[J]. 计算机工程，2007, 33(17): 46-48
Wang B, Sun L. Classification of tumor shape features based on support vector machine[J]. Computer Engineering, 2007, 33(17): 46-48

[9] 贾丹丹，李宏. 基于小波包和神经网络的血细胞识别方法的研究[J]. 中国医疗器械杂志，2008, 32(04): 239-241
Jia D D, Li H. Blood cell recognition based on wavelet packet analysis and the neural network[J]. Chinese Journal of Medical Instrumentation, 2008, 32(04): 239-241

[10] Pelikan M, *et al.* Linkage problem, distribution estimation, and Bayesian networks[J]. Evolutionary Computation, 2000, 8(3): 311-340

[11] Paragios N, *et al.* On the representation of shapes using implicit functions[J]. Statistics and Analysis of Shapes, 2006: 167-199

[12] Chan T F, Vese L A. Active contours without edges[J]. IEEE Transactions on Image Processing, 2001, 10(2): 266-277

A Bayes Classification Algorithm Based on Shape Information

WANG Yan[1,3], LIU Jun[1], CHAI A-li[2], HUANG Hai-yang[1,†],
HUAN Zhong-dan[1], LI Bao-ju[2]

(1- School of Mathematical Sciences, Beijing Normal University, Laboratory of Mathematics and Complex Systems, Ministry of Education, Beijing 100875;
2- Institute of Vegetables and Flowers, Chinese Academy of Agriculture Science, Beijing 100081;
3- Beijing Vocational College of Finance and Commerce, Beijing 101101)

Abstract: In this paper, a new Bayes classification algorithm based on the shape information is proposed to classify the objects in the image. In this method, an energy functional which indicates the similarity of different shapes is introduced to calculate the prior probability of the shape information by applying the image edge extraction and image registration algorithms, and then the objects are clustered by the Bayes method with some posterior probabilities of other features. The presented algorithm has been applied to a practical pathogeny bacteria image classification problem, and the experimental results show the high efficiency of our algorithm, which not only reduces the feature dimensions of samples, but also improves the classification accuracy. Moreover, it can fulfill the requirement of computing speed in the practical problem.

Keywords: Bayes classifier; shape information; shape similarity registration; object

Received: 05 Mar 2009. **Accepted:** 31 Dec 2009.

Foundation item: The National Natural Science Foundation of China (10531040); the National High Technology Research and Development Program of China (2006AA10Z210).

†**Corresponding author:** H. Huang. E-mail address: hhywsg@bnu.edu.cn